Srđan Radović i Aleksa Stefanović

Izdvajanje vokala iz audio zapisa

Prikazana su tri metoda za izdvajanje vokala iz audio zapisa i upoređene njihove performanse. Prvi algoritam, REPET, određuje segment u signalu koji se periodično ponavlja smatrajući ga muzikom, dok ostatak signala predstavlja vokal. Drugi algoritam je upotreba horizontalnih i vertikalnih medijanskih filtara koji na osnovu karakteristika glasa odvajaju melodijske i perkusione instrumente. Treći algoritam, PLCA, se zasniva na probabilističkom kreiranju spektrograma glasa na osnovu prethodno izdvojenih karakteristika signala koji sadrže vokale. Rezultati su pokazali da je korišćenje high-pass filtra sa graničnom vrednošću od 100 Hz dalo bolje rezultate i time je potvđena pretpostavka da retko koji ljudski glas ima frekvenciju nižu od te. Najbolje rezultate je postigao algoritam medijanskih filtara (vrednost SDR parametra za glas 5.55 dB), dok su REPET i PLCA imali slabije rezultate (2.93, odnosno 2.84 dB). U slučaju sukcesivnog korišćenja dva algoritma razlika u vrednosti SDR parametra za glas između kombinacije filtri-REPET i filtara je svega 0.14dB, što ne predstavlja veliki napredak. Što se tiče vremena obrade jednog minuta signala, REPET je pokazao najbolje performanse sa rezultatom od 0.9 s, dok su filtri (14.2 s) i PLCA (13.8 s) značajno sporiji. Ova velika razlika u vremenu predstavlja prednost REPET-a pri svakodnevnoj upotrebi. Zaključak je da izbor algoritma koji će se koristiti zavisi od potrebe korisnika. Ukoliko je ključni faktor kvalitet, preporučuju se medijanski filtri, a ukoliko je vreme najbitnije, REPET je najbolji kandidat.

Srđan Radović (1999), Bačko Dobro Polje, Vojvođanska 56, učenik 3. razreda Gimnazije „Jovan Jovanović Zmaj" u Novom Sadu

Aleksa Stefanović (2000), Pirot, Jevrejska 4/23, učenik 2. razreda Gimnazije Pirot

MENTORI:

Pavle Šoškić, student Elektrotehničkog fakulteta Univerziteta u Beogradu

Andrea Ćirić, studentkinja Elektrotehničkog fakulteta Univerziteta u Beogradu

Uvod

Izdvajanje vokala iz jednokanalnog (mono) audio snimka je u poslednjih 25 godina jedan od najzahtevnijih i najistraživanijih problema u digitalnoj obradi signala. Rešenja ovog problema imaju dosta primena u savremenom životu. Izdvojeni vokalni signal može da se upotrebi za neku dalju obradu, poput identifikacije govornika, dok instrumentalni signal ima primenu u određivanju instrumenata, transkripciji melodije ili u karaoke igrama.

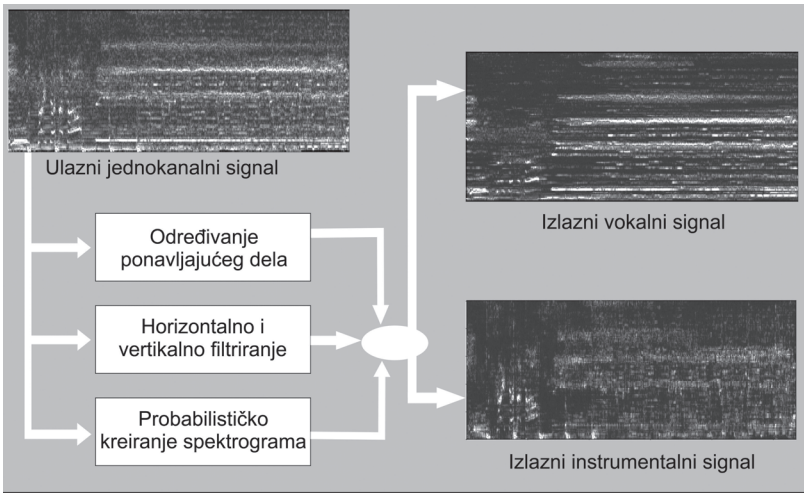
Postoji više metoda izdvajanja vokalnog signala, zasnovanih na različitim principima i pristupima problemu, a većina se bazira na obradi spektrograma samog signala, jer se sa njega može očitati najviše pouzdanih podataka pogodnih za dalju obradu (Wyse 2017). Pristupi mogu biti raznovrsni, poput probabilističkog metoda, odvajanja harmonije od perkusija, dekompozicije matrice ili izdvajanja segmenata koji se ponavljaju. U ovom radu su analizirana tri različita pristupa rešavanju ovog problema.

Prvi metod za separaciju vokala, repeating pattern extraction technique (REPET), zasniva se na izdvajanju najboljeg segmenta koji se ponavlja kroz audio zapis (Rafii i Pardo 2013). Po nemačkom naučniku Schenkeru, ponavljanje je „baza muzike kao umetnosti" (Schenker 1954). On smatra da u svakoj muzičkoj kompoziciji postoji određeni segment koji se ponavlja. U savremenijim kompozicijama, dolazi do podele pesme na strofe, prelaze, refrene, a samim tim i do promene ponavljajućeg dela. Originalni REPET algoritam je prilagođen za kraće segmente, dok su za duže segmente ili cele pesme kreirani Adaptive REPET i REPET-SIM koji kreiraju periode ponavljanja i u slučaju neizraženih delova i kada se ponavljajući delovi ne nadovezuju jedan na drugi (Rafii i Pardo 2013). U radu se ispituje uspešnost rada originalnog REPET-a i REPET-a sa high-pass filtrom.

Drugi metod koristi horizontalne, vertikalne i dijagonalne medijanske filtre. Teorijski se zasniva na tome da glas visoke frekvencije traje kratko i da je sličan perkusijama, dok je u nižoj frekvenciji duži i sličan melodiji, tj. instrumentima koji prave melodiju, poput klavira ili gitare. Horizontalni filtri služe da se u vremenskom periodu smanji pojava perkusija, dok vertikalni u određenom frekvencijskom opsegu uklanjaju horizontalne komponente. Kao dodatak originalnom algoritmu sa horizontalnim i vertikalnim filtrima, koriste se dijagonalni filtri koji imaju zadatak da sačuvaju glasovne signale koji su na spektrogramu predstavljeni dijagonalama (Deif *et al*. 2015). Upoređuju se rezultati korišćenja horizontalnih i vertikalnih filtara i korišćenja dijagonalnih filtara kao dodatka.

Treći metod, probabilistic latent component analysis (PLCA), koristi probabilističke metode u određivanju glasovnog signala. Sastoji se iz dva dela. U prvom delu koristi support vector machine (SVM) da bi označio pozadinu u ulaznom signalu. Drugi deo je expectation maximization (EM) algoritam i on uči kako da napravi spektrogram identičan ulaznom. Koristi latentne promenljive, nasumično odabrane na početku, da bi izrazio verovatnoću pojave određene frekvencije u određenom trenutku u vremenu. Na osnovu odnosa sa ulaznim spektrogramom optimizuju se vrednosti latentnih promenljivih i većim brojem iteracija se dolazi do idealnog spektrograma. U prvom krugu se dobija idealna pozadina, a u drugom se izdvaja glas iz signala glasa i pozadine (Mendez *et al*. 2012).

Cilj ovog rada je da uporedi rad i rezultate korišćenjem tri prethodno navedena algoritma. Hipoteze su da će korišćenje high-pass filtara sa tresholdom 100 Hz poboljšati rezultate svih algoritama, da će dijagonalni filtri doprineti boljim rezultatima prilikom korišćenih filtara, kao i da će sukcesivno korišćenje dva algoritma dati bolje rezultate od pojedinačnog korišćenja algoritama. Analizira se i vreme izvršavanja algoritama.



Slika 1.
Blok dijagram celokupnog sistema: Ulazni signal se obrađuje korišćenjem jednog ili više algoritama i na izlazu se dobijaju dva signala, vokalni i instrumentalni.

Figure 1. Block diagram of the whole system: Input signal is being processed using one or more of the algorithms and on the output two signals are given, vocal and instrumental.

Metod

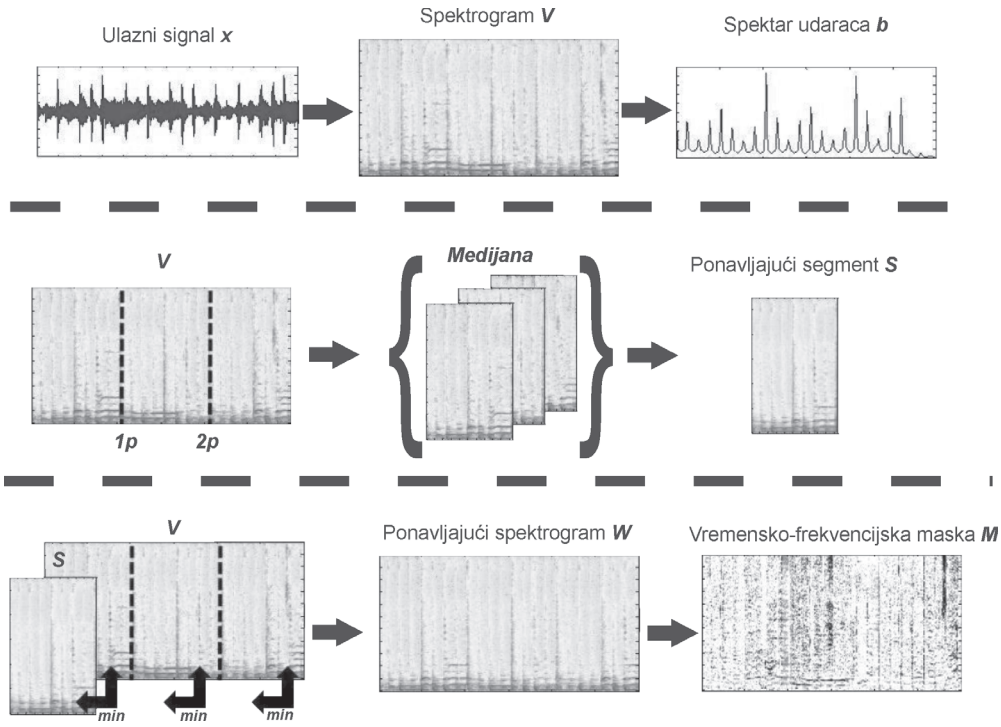
REPET (REpeating Pattern Extraction Technique)

REPET je algoritam koji analizom spektrograma ulaznog signala detektuje određene delove koji se nakon nekog perioda ponavljaju. Ti delovi koji su periodično ponavljajući se smatraju muzikom, dok ostali, neponavljajući deo, pripada glasu. Sam algoritam možemo podeliti u 3 faze: nalaženje najboljeg perioda ponavljanja, određivanje segmenta koji se ponavlja i samo izdvajanje ponavljajućeg i neponavljajućeg dela signala (Rafii i Pardo 2013)

Određivanje perioda ponavljanja. Jedan od načina da se odredi period ponavljanja jeste metod autokorelacije. Metod se zasniva na određivanju sličnosti određenog segmenta sa njegovim kasnijim pojavljivanjem.

Ulazni podatak ovog algoritma je jednokanalni audio signal x. Na početku je, u svrhu dalje obrade signala, neophodno da se Short-time Furijeovom transformacijom (STFT) signal iz amplitudnog domena prevede u frekvencijski domen. Određivanjem apsolutnih vrednosti svih elemenata prethodno dobijenog rezultata dobija se spektrogram ulaznog signala V. Kvadriranjem spektrograma V povećavaju se razlike među frekvencijama i smanjuju intenziteti nižih frekvencija, koje bi mogle predstavljati šumove u signalu. Nakon dobijenog spektrograma V^2 vrši se autokorelacija svakog reda spektrograma kako bi se uočile sličnosti između susednih segmenata. Novodobijena matrica B se zatim usrednjuje i dobijamo takozvani spektar udaraca b. Radi lakšeg upoređivanja, rezultati b se normalizuju sa početnim elementom. Ceo proces je opisan sledećim jednačinama:

$$B(i,j) = \frac{1}{m-j+1} \sum_{k=1}^{m-j+1} V(i,k)^2 V(i,k+j-1)^2,$$



Slika 2. Blok dijagram REPET algoritma

Figure 2. Block diagram of REPET algorithm

$$b(j) = \frac{1}{n}\sum_{i=1}^{n} B(i, j).$$

$$b(j) = \frac{b(j)}{b(1)}.$$

gde se i menja od 1 do $(N + 1)/2$ (N – broj frekvencijskih opsega), a j od 1 do m (broj vremenskih opsega). Dobijeni spektar udaraca prikazan je na slici 3. Sa slike se primećuje da postoje određeni elementi izrazito velike vrednosti. Oni predstavljaju početak jednog takta. Periodi se mogu odrediti kao rastojanja između dva elementa velikih vrednosti. Pošto postoji više mogućih perioda, potrebno je odrediti onaj najbolji, tj. onaj koji ima najveću srednju akumuliranu energiju nad svojim višestrukim vrednostima. Za svaki mogući period i i njegove celobrojne umnoške $2i$, $3i$, $4i$, … proverava se da li se ektremne vrednosti poklopaju sa pozicijom perioda. Pošto nekad neće doći do preklapanja, opseg za proveravanje poklapanja se proširuje na $[i - \Delta, i + \Delta]$, gde je Δ varijabla kojom se određuje opseg greške, u ovom slučaju vrednost varijable je $j/2$ (Rafii i Pardo 2013). Ukoliko se periodi poklapaju sa ekstremnim vrednostima oni se sabiraju a zatim se oduzima

srednja vrednost svih vrednosti u opsegu greške (do Δ), koja se smatra određenim šumom. Nakon provere svih perioda, najviša vrednost od svih suma određuje da je njen period p najbolji (slika 2, prvi red).

Modelovanje ponavljajućeg segmenta. Nakon određenog perioda, potrebno je odrediti deo spektrograma koji se ponavlja i smatrati ga segmentom pozadine *S*. Prethodno kreirani spektrogram *V* se podeli u *r* delova sa periodom *p*. Podela se izvršavanja određivanjem medijane svakog vremenskog segmenta u signalu od 0 do *p* u svakom od delova:

$$S(i, l) = \underset{k=1 \ldots r}{\text{median}} \{V(i, l + (k-1)p)\},$$

gde se *i* menja od 1 do *n* (broj frekvencija), *l* od 1 do *m* (vreme), *k* od 1 do *r* (broj segmenata) (slika 2, drugi red).

Izdvajanje vokalnog i instrumentalnog signala. Sada kada je dobijen ponavljajući segment *S*, on se koristi za kreiranje ponavljajućeg spektrograma *W*. Spektrogram *W* se dobija uzimanjem minimalne vrednosti između spektrograma *S* i *V*, što znači da će svi delovi koji se ponavljaju činiti spektrogram *W*:

$$W(i, l + (k-1)p) = \min\{S(i, l), V(i, l + (k-1)p)\},$$

i se menja od 1 do *n* (broj frekvencija), *l* od 1 do *m* (vreme), *k* od 1 do *r* (broj segmenata). Da bi se izdvojio instrumentalni signal, potrebno je kreirati masku koja će predstavljati koji delovi ulaznog spektrograma će biti deo instrumentala. Maska se dobija pomoću sledeće jednačine:

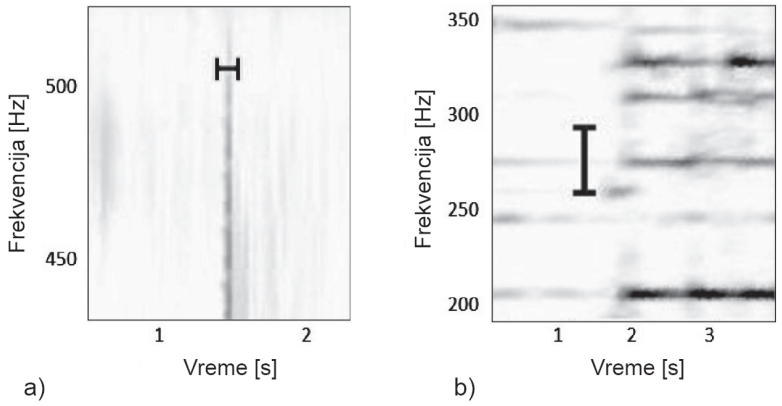
$$M(i, j) = \frac{W(i, j)}{V(i, j)}.$$

Maska se primenjuje na spektrogram *X* i dobija se instrumentalni spektrogram. Vokalni spektrogram se kreira jednostavnim oduzimanjem vrednosti instrumentalnog spektrograma od *X* (slika 2, treći red). Na vokalni spektrogram se primenjuje high-pass filtar sa graničnom vrednošću od 100 Hz, koji će sve vrednosti u vokalnom spektrogramu čije su frekvencijske vrednosti veće od 100 Hz zadržati, dok će niže pridodati instrumentalnom spektrogramu. Razlog zašto korišćenje filtra može poboljšati rezultate je zato što retko koji ljudski glas ima frekvenciju nižu od 100 Hz, kao i što određeni deo signala niže frekvencije predstavlja šum signala.

Medijanski filtri

Glavna ideja kod ovog algoritma (Deif *et al*. 2015) je odvajanje melodijskog i perkusionog dela pesme. Vokali izgledaju kao perkusije na spektrogramu visoke, dok su slični melodijskim instrumentima na spektrogramu niske frekvencijske rezolucije. Zbog toga postoje dva dela algoritma, koji rade na različitim frekvencijskim rezolucijama, da bi se vokali odvojili i od melodijskih i od perkusionih instrumenata.

Pošto se perkusije na spektrogramu ulaznog signala pojavljuju kao vertikalne linije, a melodijski instrumenti kao horizontalne, za njihovo odvajanje mogu se koristiti horizontalni, odnosno vertikalni medijanski filtri (slika 3). Medijanski filtri za vrednost datog elementa uzimaju



Slika 3.
(a) Horizontalni i
(b) vertikalni medijanski filtar

Figure 3.
(a) Horizontal and
(b) vertical median filter

medijanu elemenata iz njegove okoline (po jednoj osi), i tako vertikalni filtri uklanjaju horizontalne linije i obrnuto, ako je korišćena okolina datog elementa velika u odnosu na debljinu linije. Formule kojima se definišu filtri su:

$$H = \mathrm{MD}_{\mathrm{h}}\{V, l_{\mathrm{h}}\},$$

$$P = \mathrm{MD}_{\mathrm{v}}\{V, l_{\mathrm{v}}\},$$

gde su MD_{h} i MD_{v} medijanski filtri, V ulazni spektrogram, a l_{h} i l_{v} dužine perioda koji se uzima u obzir za određivanje medijane. U slučaju horizontalnih filtara, l_{v} iznosi 0.15 sekundi, dok l_{v} pri korišćenju vertikalnih filtara 20 Hz. Za odvajanje vokala, prvo se primene filtri na spektrogram ulaznog signala, visoke frekvencijske rezolucije, nakon čega se dobije signal melodijskih instrumenata, kao i signal kombinacije vokala i perkusionih instrumenata. Zatim se primene filtri na spektrogram kombinacije, niske frekvencijske rezolucije, i dobiju se odvojeni signali vokala i perkusija. Signal perkusija se sabere sa ranije dobijenim signalom melodijskih instrumenata da bi se dobio kompletan signal muzike.

Ovaj algoritam se može poboljšati dodatnim korišćenjem dijagonalnih medijanskih filtara prilikom odvajanja vokala od perkusija. Primenjuje se odvojeno 6 dijagonalnih filtara, sa uglovima od ±30°, ±45° i ±60°. Zatim se kreira zajednički spektrogram za horizontalne i dijagonalne filtere na sledeći način:

$$H' = \max(H, D_1, D_2, \ldots, D_6).$$

Razlog za uvođenje dijagonalnih filtara je pojava i dijagonalnih linija u spektrogramu vokala, koje se ne očuvaju pri primeni samo horizontalnih i vertikalnih filtara.

PLCA (Probabilistic Latent Component Analysis)

Ovaj algoritam (Mendez *et al*. 2012) je zasnovan na mašinskom učenju, što znači da, za razliku od prošla dva algoritma, ne koristi nikakve unapred određene karakteristike vokala ili muzike, ponavljanje kod REPET-a ili frekvencijske karakteristike glasa kod medijanskih filtara.

Pored ulaznog signala, ovom algoritmu treba proslediti i segmente tog signala koji ne sadrže vokale već samo muziku. Na tim segmentima algoritam nauči šta je muzika u datom signalu i može je odvojiti. Izabrano je da se koriste segmenti bez vokala zato što su oni mnogo češći od segmenata koji sadrže vokale ali ne i muziku.

Da bi se automatizovao proces određivanja segmenata bez vokala, koristi se algoritam za klasifikaciju, SVM. SVM se jednom istrenira koristeći segmente nekih pesama za koje znamo da ne sadrže vokale, kao i segmente za koje znamo da sadrže vokale. Za treniranje se ne koriste segmenti kao sirov audio signal, već se izvuku neke osobine: spektralna ravnost (koliko je šumovit signal), spektralni centroid (centar mase spektrograma), root-mean-squared amplituda spektra, i 25 mel koeficijenata. Nakon treniranja, SVM za svaku datu pesmu određuje segmente bez vokala i prosleđuje ih u PLCA. Kernel koji je korišćen je kvadratni, jer je davao bolje rezultate od linearnog i Gausovog.

Sam PLCA je Expectation-Maximization algoritam. Takvi algoritmi sadrže određeni broj latentnih promenljivih, koje na početku dobijaju nasumične vrednosti, a kasnije se menjaju kroz veliki broj iteracija i postaju sve bliže željenom obliku. PLCA se sastoji iz dva dela. U prvom delu latentne promenljive postaju sve sličnije segmentima bez vokala, i tako se dobiju vrednosti latentnih promenljivih za muziku. U drugom delu se dobijaju latentne promenljive čitavog ulaznog signala, ali tako što se koriste prethodno dobijene latentne promenljive za muziku i njima se dodaju nove latentne promenljive koje predstavljaju vokale. Tokom iteracija, latentne promenljive za muziku se ne menjaju, već samo one za glas. Tako se nakon završetka drugog dela dobijaju odvojeno latentne promenljive za glas i muziku, koje se pretvaraju u signal glasa i muzike.

Korišćenje dva algoritma

Jedan od načina za poboljšanje kvaliteta izlaznog signala jeste sukcesivno korišćenje dva algoritma. Ovaj metod se zasniva na tome da izlazni vokalni signal prvog algoritma postane ulazni signal drugog koji za zadatak ima da dodatnom obradom odstrani eventualne elemente muzike ili novonastale šumove i da se izlazni vokalni iskoristi za evaluaciju. Nakon dodatne obrade vokalnog signala, sledi obrada instrumentalnog signala takođe kroz drugi algoritam. Pored kvaliteta izlaznog signala, povećava se i vreme izvršavanja zato što se drugi algoritam izvršava dva puta. To može predstavljati problem u svakodnevnoj upotrebi ukoliko je vreme presudni faktor.

Rezultati

Baza podataka. Za testiranje performansi algoritama, korišćena je MIR-1K baza podataka. Sastoji se iz 1000 stereo audio snimaka dužine 5 do 13 sekundi. Na levom kanalu je instrumentalni deo određene pesme, dok je na desnom kanalu glas. Snimci predstavljaju delove 55 izvođenja kineskih

karaoke pesama. Na početku rada svakog algoritma, od stereo snimaka kreiran je jednokanalni snimak kako bi izdvajanje bilo uspešno.

Evaluacija. Evaluacija rada algoritama se zasniva na upoređivanju dobijenih signala sa početnim. Izdvojeni vokalni signal se upoređuje sa desnim kanalom ulaznog stereo signala, dok se izdvojeni instrumentalni signal upoređuje sa levim kanalom ulaznog signala. Za evaluaciju je korišćena biblioteka BSS_EVAL toolbox (Févotte *et al*. 2005). Na početku je potrebno podeliti spektrogram izdvojenog signala na komponente koje su pogodne za upoređivanje na sledeći način:

$$s_{\text{out}}(t) = s_{\text{interf}}(t) + e_{\text{artif}}(t),$$

gde je s_{out} spektrogram izlaznog signala, s_{e} dozvoljena distorzija izvora sa unapred određenim vrednostima, e_{interf} predstavlja dozvoljene promene u signalu vezane za pojavu neželjenih izvora, a e_{artif} se odnosi na pojavu artefakata, koji su u vidu šuma ili grešaka koje su nastale tokom algoritamske obrade.

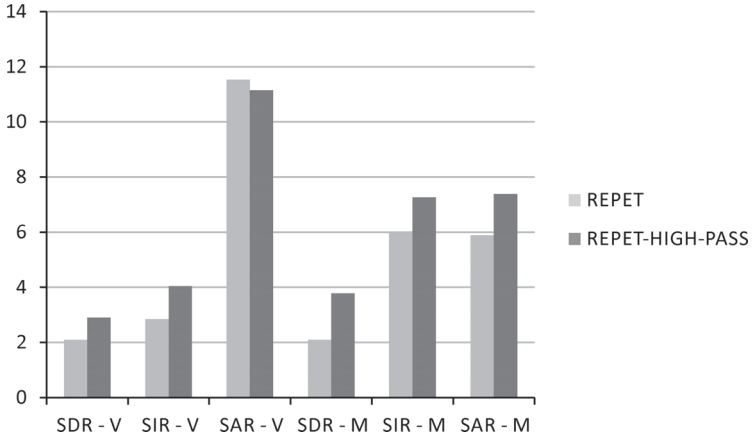
Veličine koje se koriste za opisivanje kvaliteta izdvojenog signala su izražene sledećim formulama:

$$\text{SDR} = 10\log_{10}\left(\frac{\left\| s_{\text{e}} \right\|^2}{\left\| e_{\text{interf}} + e_{\text{artif}} \right\|^2}\right),$$

$$\text{SIR} = 10\log_{10}\left(\frac{\left\| s_{\text{e}} \right\|^2}{\left\| e_{\text{interf}} \right\|^2}\right),$$

$$\text{SAR} = 10\log_{10}\left(\frac{\left\| s_{\text{e}} + e_{\text{interf}} \right\|^2}{\left\| e_{\text{artif}} \right\|^2}\right),$$

gde SDR (Source-to-Distortion-Ratio) predstavlja ukupno odsustvo distorzije u signalu i sličnost između izdvojenog i ulaznog signala, SIR (Source--to-Interferences-Ratio) označava prisustvo instrumentalnog dela signala u vokalnom i obrnuto, SAR (Source-to-Artifacts-Ratio) predstavlja zastu-



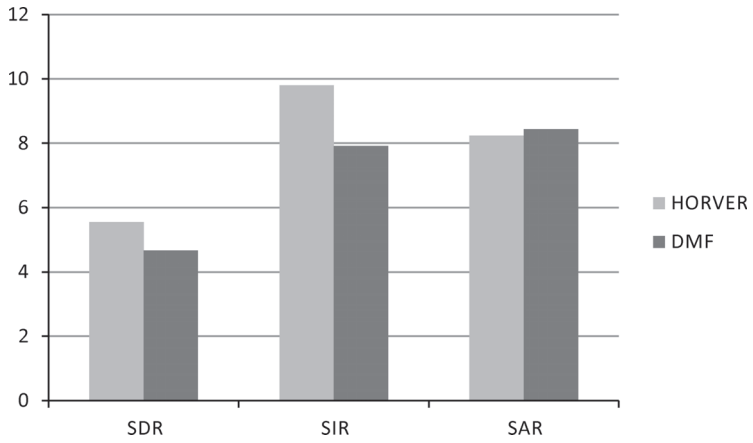
Slika 4.
Zavisnost vrednosti parametara od korišćenja high-pass filtra

Figure 4.
Dependence of values of parameters by usage of high-pass filter

pljenost artefakata. Za sve veličine važi da veće vrednosti znače bolji kvalitet izdvojenog signala.

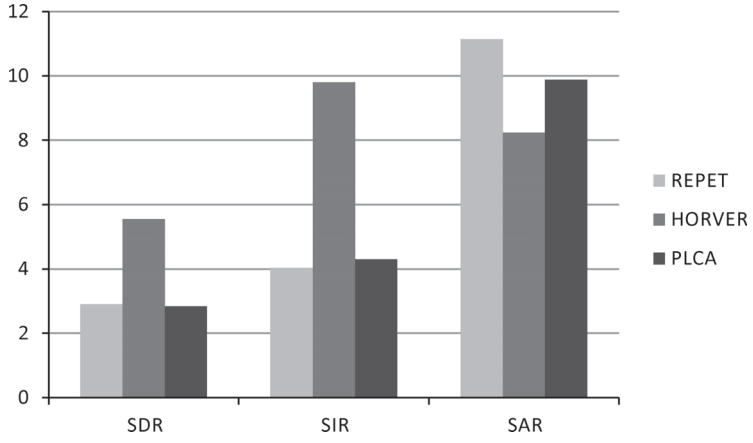
Sa slike 4 se može primetiti da vrednosti 5 od 6 parametara imaju veće vrednosti ako se koristi high-pass filtra sa graničnom vrednošću od 100 Hz. Korišćenjem filtra se javlja više artefakata u vokalnom signalu, dok je zastupljeno manje muzike u vokalima i obrnuto. Na osnovu rezultata pri korišćenju REPET algoritma, high-pass filtar je primenjen i na ostala dva algoritma.

Rezultati sa slike 5 pokazuju da korišćenje dijagonalnih filtara pored horizontalnih i vertikalnih ne daje bolje rezultate. Količina artefakata je umanjena za 0.2 dB, što ne predstavlja značajno poboljšanje. Zbog određene zastupljenosti dijagonalnih komponenti i kod instrumenata, povećano je prisustvo instrumentalnog signala u vokalnom.



Slika 5.
Zavisnost vrednosti parametara za glas od korišćenja dijagonalnih filtara

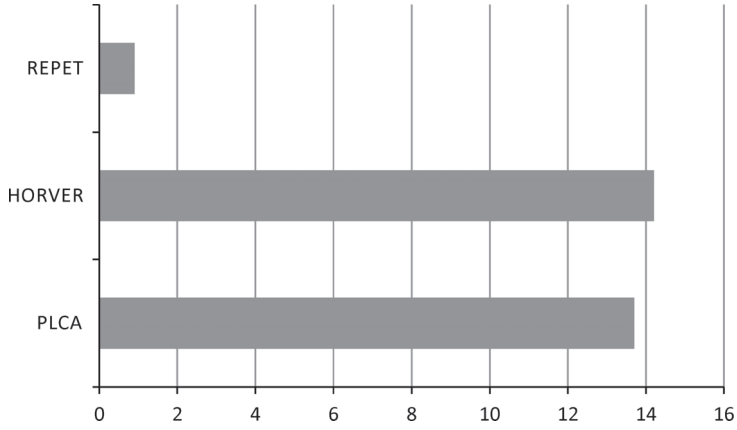
Figure 5.
Dependence of values of parameters for vocals by usage of diagonal filters



Slika 6.
Zavisnost vrednosti parametara za glas od korišćenog algoritma

Figure 6.
Dependence of values of parameters for vocals by used algorithm

Sa slike 6 se primećuje da korišćenje medijanskih filtara daje znatno bolje rezultate za SDR i SIR parametre od REPET-a i PLCA, koji imaju približno iste rezultate. Razlog za veliku razliku u rezultatima je što filtri najtačnije određuju razlike između glasa i instrumenata, dok REPET za-



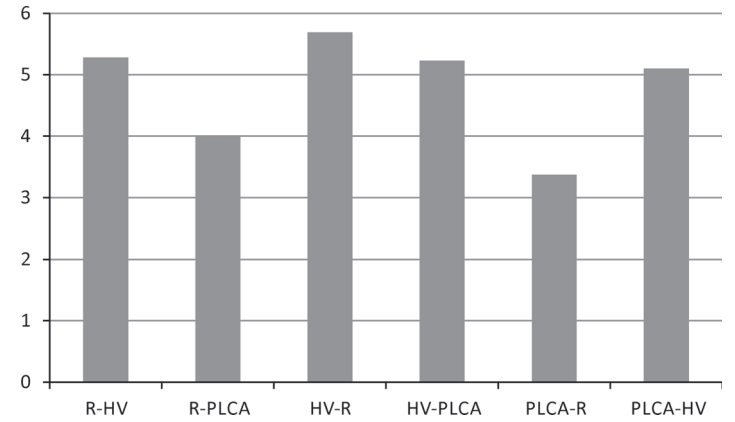
Slika 7.
Zavisnost vremena izvršavanja od korišćenog algoritma

Figure 7.
Dependence of executing time by used algorithm

stupa teoriju neperiodičnosti glasa, koja često nije potpuno tačna, a PLCA daje loše rezultate zbog malog broja signala u trening setu. Ukoliko se PLCA algoritmu proslede već obeleženi signali (ukoliko se ne primenjuje SVM), rezultati su u rangu medijanskih filtara uz grešku od 0.1 dB. Mana medijanskih filtara je što se javlja dosta artefakata, u čemu su ostala dva algoritma znatno bolja.

Vreme izvršavanja REPET algoritma je 0.9 s, dok je u slučaju filtara i PLCA vreme 14.2 odnosno 13.8 sekundi (slika 7), mereno na procesoru Intel(R) Core(TM) i7-2640M. Značajna razlika u vremenima je zbog kompleksnosti algoritama, zbog čega je REPET znatno primenljiviji za svakodnevnu upotrebu.

Pri sukcesivnom korišćenju dva algoritma, najbolje rezultate dala je kombinacija filtri–REPET (slika 8). Može se primetiti da su najbolji rezultati oni u kojima su filtri na prvom mestu, što direktno sledi iz činjenice da filtri samostalno daju najbolje rezultate, pa samim tim drugi algoritam ima zadatak da popravi kvalitet signala. Pri upoređivanju rezultata sa pojedinačnim algoritmima, kombinacija filtara i REPET-a daje bolje rezultate u odnosu na filtre za 0.14 dB (5.69 u odnosu na 5.55 dB). Ova razlika ne predstavlja značajno poboljšanje kvaliteta, ali povećava vreme izvršavanja, tako da je primenljivost u praksi znatno manja.



Slika 8.
Zavisnost vrednosti SDR parametra za glas od korišćene kombinacije algoritama

Figure 8.
Dependence of value of SDR parameter for voice by used combination of algorithms

Zaključak

Cilj ovog rada je bio da se uporede performanse tri algoritma za izdvajanje vokala iz audio signala koji se zasnivaju na tri različite teorije: ponavljućim segmentima, horizontalnim i vertikalnim karakteristikama glasa i probabilističkom kreiranju spektrograma. U kontekstu kvaliteta izdvojenog signala, najbolje rezultate je postigao algoritam medijanskih filtara zbog najtemeljnije obrade karakteristika glasa. Potvrđena je hipoteza da će korišćenje high-pass filtra dati bolje rezultate, i opovrgnuta druga hipoteza vezana za korišćenje dijagonalnih filtara, koji su pored dijagonalnih komponenti glasa očuvali i dijagonalne komponente muzike. Što se tiče vremena izvršavanja, REPET je davao višestruko bolje rezultate, što može značiti da je znatno primenljiviji u praksi od ostala dva. Pokazalo se da korišćenje dva algoritma ne doprinosi značajnijem poboljšanju kvaliteta, ali doprinosi vremenu izvršavanja.

Glavni zaključak je taj da, u zavisnosti od potreba korisnika, on sam bira koji će algoritam koristiti. Ukoliko je kvalitet signala ključan, preporučuje se korišćenje medijanskih filtara. Ukoliko je vreme najbitnije, REPET može biti primenljiviji i pored slabijih performansi. PLCA u bilo kom slučaju nije najbolje rešenje i ne preporučuje se njegovo korišćenje. Dalji rad na ovu temu može biti kreiranje Android aplikacije koja će rad ovog projekta prikazati široj javnosti.

Literatura

Deif H., Fitzgerald D., Wang W., Gan L. 2015. Separation of Vocals From Monaural Music Recordings Using Diagonal Median Filters and Practical Time-Frequency Parameters. U *2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, str. 93-97.

Févotte C., Gribonval R., Vincent E. 2005. BSS_EVAL Toolbox User Guide – Revision 2.0.

Mendez D., Pondicherry T., Young C. 2012. Extracting vocal sources from master audio recordings. http://cs229.stanford.edu/proj2012/MendezPondicherryYoung-ExtractingVocalSourcesFromMasterAudioRecordings.pdf

Rafii Z., Pardo B. 2013. REpeating Pattern Extraction Technique (REPET): A simple method for music/voice separation. *IEEE Transactions on Audio, Speech, and Language Processing*, **21** (1): 71.

Schenker H. 1954. *Harmony*. Chicago: University of Chicago Press

Wyse L. 2017. Audio Spectrogram Representations for Processing with Convolutional Neural Networks. U *Proceedings of the First International Workshop on Deep Learning and Music joint with IJCNN*. Anchorage, US, May, 2017: str. 37-41.

Srđan Radović and Aleksa Stefanović

Extracting Vocals from Songs

In this paper three methods for voice extraction from audio signals are presented and their performances are compared. The first algorithm, REPET, detects a segment in the signal which is periodically repeating considering it music, while the rest of the signal represents vocals. The second algorithm is usage of horizontal and vertical median filters which, based on characteristics of voice, separate melodic and percussive instruments. An addition for this algorithm is the usage of diagonal filters, which serve to save diagonal components of vocals. Results show that the diagonal filters were not a good solution for improving quality. The third algorithm, PLCA, is based on the probabilistic creating of a voice spectrogram by analyzing previously extracted characteristics of signals which contain vocals. Results have shown that the usage of a high-pass filter with a threshold of 100 Hz gave better results and with this an assumption that rarely any human voice has frequency below that one is proved. The best results were made by the median filter algorithm: the value of the SDR parameter (source to distortion ratio) for voice was 5.55 dB, while REPET and PLCA had worse results (2.93 dB and 2.84 dB). In the case of successive usage of two algorithms, which was to significantly improve results because of more detailed signal processing, the difference of value of SDR for voice between the combination filters-REPET and filters is just 0.14 dB, which is not a big improvement. Concerning the time of processing of one minute of signal, REPET has shown best performances with a time of 0.9 s, while filters (14.2 s) and PLCA (13.8 s) were significantly slower. This big difference between times is an advantage of REPET in everyday use. The main conclusion is that the choice of algorithm which will be used depends on the user's needs. If the key factor is quality, median filters are recommended, and if time is the most important, REPET is the best candidate.